Word order variations and spoken man-machine dialogue in French : a corpus analysis on the ATIS domain

Jean-Yves Antoine, Jérôme Goulian

VALORIA, University of South Britanny
IUP Vannes, rue Yves Mainguy, F-56000 Vannes, France
Phone : +33 2 97 68 32 10 — Email : {Jean-Yves.Antoine,Jerome.Goulian}@univ-ubs.fr
WWW : http://www.univ-ubs.fr/valoria/antoine

1. Introduction

During the last decade, spoken man-machine dialogue has known significant improvements that should lead shortly to the development of real use systems. In spite of these indisputable advances, numerous limitations restrict still the expansion of common use spoken dialogue systems. In particular, present researches in spoken man-machine communication lack seriously genericity. Most of spoken dialogue systems are indeed concerned by a unique application domain : transport information (ATIS domain). This task is very restricted, what allows the achievement of *ad hoc* processing methods that ignore most of the structure of the sentence — see for instance (Minker, 1999) for an overview concerning speech understanding. Although these approaches show a significant robustness on spontaneous speech, their portability to other application domains remains an open issue (Hirschman, 1998) : one should reasonably assume that less restricted tasks require a more detailed linguistic analysis.

As a result, future advances of man-machine communication depend on the improvement of current spoken language models. In our opinion, corpus linguistics should be of great help for the development of such improved models :

- the analysis of large task-specific corpora should provide a precise characterisation of the linguistic phenomena that occur in the concerned application domain. This characterisation is very useful for the achievement of a system prototype[1], but should be helpful for evaluation purposes too (Antoine et al, 1999).
- the comparison of different corpora should assess usefully the linguistic variabilities — and their causes : task influence, familiarisation with the task, kind of user,... — that should occur in spoken man-machine dialogue. It should therefore provide answers to the important problem of genericity.

In this paper, we present a corpus analysis which concerns specifically word order variations in spoken French. This analysis has been carried out on a corpus of spoken man-man dialogue — the Air France corpus (Morel *et al.*, 1989) — that corresponds to the ATIS domain. At first, this paper presents briefly the problem of word order freedom and its implications for natural language processing. We then detail the main results of this corpus analysis from a strictly linguistic point of view. In particular, we assess the influence of the familiarisation with the task by means of a comparison of two subparts of the corpus (see section 3). We finally discuss the consequences of these linguistic observations on the achievement of spoken dialogue systems.

2. Word order and natural language processing

Word order is an important question for human language technologies. For instance, the problem of word order freedom originated the development of dependency grammars (Tesnière 1990, Mel'cuk 1988) in response to some weaknesses of standard phrase structure grammars[2].

[1] Several works have shown for instance that the errors of probabilistic language models may be the result of systematic failures on a restricted number of linguistic phenomena.

[2] This controversy is stille open : see for instance (Rambow and Joshi, 1994) or (Pollard and Sag, 1994).

Likewise, stochastic language models (N-grams) depend to a large extent on word order. As a result, any increase of word order variations should increase harmfully the perplexity of the language model.

Generally speaking, two kinds of word order freedom should be distinguished (Holan *et al*., 2000) :

- *weak word order freedom* — called *freedom of constituent order within a continuous head domain* by Holan — where a constituent is free to move in several places but remains always continuous. The corresponding utterance respects therefore the constraint of projectivity. For instance :

  (1) *on the morning Paul used to go shopping*.

- *global word order freedom* which corresponds on the opposite to a relaxation of continuity. In such cases, some extracted elements are allowed to appear out of the constituent they are supposed to belong to : the corresponding utterance is therefore non projective. Consider for instance the following *wh*-extraction (Hudson, 2000) :

  (2) *who do you think that Mary claims that Sarah likes*

Global word order freedom concerns above all free word order languages (Russian, Finnish, Czech,...) whereas rigid languages (English for instance) are more concerned by weak variability (Holan *et al.*, 2000). Written French should be considered as a rigid word order language (Covington, 1990). Spoken French is however hardly identifiable to written French (Blanche-Benveniste *et al*., 1990). There is therefore no linguistic evidence that spontaneous spoken French presents a word order variability that is only restricted to weak variation.

Besides the important question of the processing of non projective structures (global word order freedom), weak variability constitutes a not inconsiderable problem for spoken language technologies. Since speech recognition provides usually several hypothetical utterances (*N-best sequences*), any increase of ambiguity / perplexity due to weak variability should affect noticeably the robustness of the system. The variability of spontaneous spoken French is therefore an important problem from a computational point of view. The corpus analysis detailed in this paper aims precisely at answering this question on a specific task domain.

3. Air France corpus

This analysis has been carried out on a speech corpus which was transcribed from the recording of real dialogues between a hostess of an air transport information service (ATIS domain) and several customers (Air France corpus). This corpus represents 103 dialogues that correspond to 5149 speech turns and 49703 words. It has been divided into two corpora which correspond respectively to individual customers and travel agents (figure 1) in order to assess the influence of the familiarisation with the task.

Table 1 — Description of the Air France corpus.

Corpus	number of dialogues	number of speech turns	number of words	familiarisation with the task
individual customers	68	3676	n.c.	low
travel agents	35	1473	n.c.	high
Total	103	5149	49703	—

This corpus does not correspond to a man-machine interaction, but on the contrary to a dialogue between two humans. As a matter of fact, our purpose is to characterise the real usages that should be modelled by spoken dialogue systems.

4. Corpus analysis

We have made an exhaustive inventory of all the extractions that occur in the Air France corpus. Every observed phenomenon has been characterised according to several features (Gadet, 1992).

1) *direction of the extraction* — anteposition (movement of an element to the left of the utterance) or postposition (movement to the right),

2) *kind of construction* — we have distinguished the following kind of extracted constructions :

- simple inversion (extraction without any specific linguistic mark) :

sur Héraklion on n'a qu'un seul tarif special (AF.II.17.O14)

[*for Heraklion* we have only one unique special fare]

- dislocations (extraction marked by a clitic)

le visa on l'a eu au consulat (AF.I.48.C6)
[*the visa* we have obtained *it* at the consulate]

- presentative structures (among which cleaved sentences)

j'ai quelqu'un qu'est allé prendre des billets charters pour moi (AF.I.43.C9)
[*I have* (there is) *someone that* went and took charter tickets for me]

c'est une personne de nationalité tunisienne qui a eu ce billet (AF.I.4.O7)
[*This is* a Tunisian *that* had this ticket]

3) *syntactic function of the extracted element* — subject, argument, adjunct or finally sentence complement — also called *associés* (associated elements) in (Blanche-Benveniste, 1997).

4) *projectivity* — continuous or discontinuous extraction.

These linguistic features have been characterised by their frequencies of occurrence in the Air France corpus. Since the notion of sentence is not relevant in spoken French (Blanche-Benveniste *et al.*, 1990), speech turns has been used as unit of segmentation for the computation of these probabilities.

5. Quantitative importance of word order variations in spontaneous spoken French

At first, we present some conclusions that can be drawn from the observation of the whole corpus. The influence of the familiarisation with the task will be discussed in the following section, which concerns the comparison between the two sub-corpora (individual customers and travel agents).

First of all, spontaneous spoken French seems to be — given the considered task — noticeably more flexible than written French. Table 2 shows[3] indeed that a not inconsiderable part (13.6%) of the speech turns presents at least one word order variation.

Table 2 — Frequency of word order variations in the Air France corpus (mean number of speech turns presenting at least one extraction).

Corpus	mean frequency	standard variation	minimal frequency on a dialogue	maximal frequency on a dialogue
individual customers	14.9%	6.9%	0.0 %	29.8 %
travel agents	10.1%	8.2%	0.0 %	30.8 %
Total	**13.6 %**	**7.5 %**	**0.0 %**	**30.8 %**

Besides, the statistical distribution of these frequencies presents a high standard deviation. The use of word order variations is therefore very variable from a dialogue to an other. It is quite difficult to explain this variability by means of a unique cause. For instance, dialogic context (negociation, reformulating,...) is undoubtedly a noticeable source of variability, but one might reasonably assume that idiosyncratic factors can intervene too. Anyway, it appears that word order variations are rather common in spoken French. This is why they can not be ignored in the prospect of a robust spoken

[3] This table and all the following ones present global results computed on the whole corpus as well as particular results observed on the sub-corpora that concern respectively "individual customers" and "travel agents". The comparison of these last two corpora will be discussed in section 8.

man-machine communication. Fortunately, a detailed analysis of the observed extractions shows that the latter respect to a certain extent some rigid word order constraints.

6. Constrained extractions : projectivity and SVO canonical order

As shown by table 3, most of word order variations correspond unsurprisingly (Gadet, 1992) to antepositions (82.5 % of the observed variations). This difference between antepositions and postpositions is statistically significant (χ^2 test[4] of an identical distribution : $CHI_{AF} = 0.997$).

Table 3 — distribution of the extractions according to their direction (mean percentage of antepositions and postpositions).

Corpus	anteposition	postposition	standard deviation
individual customers	82.9%	17.1%	18.4%
travel agents	81.2%	18.8%	24.1%
Total	**82.5 %**	**17.5 %**	**20.4 %**

This predominance of the antepositions can be related to the distribution of the extracted elements according to their syntactic function (table 4). Word order variations concern above all subjects (30.7 % of the observed variations), sentence complements (30.0 %), adjuncts (27.4%), whereas subcategorized arguments represent only 12.0 % of the observed variations. This lesser occurrence of argument extractions is statistically significant (Student mean test[5] of identical distribution of subject and arguments : Tsub/arg = 3.652 ; T(0.01) = 2.600). On the contrary, there is no significant difference between the three other functions (Student mean test : Tsub/adj = 0.911 ; Tsub/scpl = 1.059 ; T(0.1) = 1.652).

Table 4 — distribution of the word order variations according to the syntactic function of the extracted element.

Corpus		subjects	arguments	adjuncts	sentence complements
individual customers	Mean *(St. Dev.)*	29.6% *(26.0%)*	12.6% *(15,2 %[6]).*	28.8% *(24.4%)*	29.0% *(22.4%)*
travel agents	Mean *(St. Dev.)*	34.6% *(35.6%)*	9.4% *(16,1 %)*	22.5% *(30.2%)*	33.5% *(28.3%)*
Total	Mean *(St. Dev).*	**30.7 %** *(29.6 %)*	**12.0 %** *(15.5 %)*	**27.4 %** *(26.5 %)*	**30.0 %** *(24.5 %)*

This distribution seems to be coherent from a linguistic point of view. Generally speaking, written French follows a canonical SVO (subject-verb-object) order. Since adjuncts or sentence complements are not concerned by this ordering constraint, they are relatively free to move inside the sentence. Likewise, subject extractions follow in most cases a SVO order since they correspond very frequently to an anteposition (table 5). Subject extraction is consequently rather free. On the opposite, the position of arguments is rigidly fixed by the SVO canonical order. Argument extractions are thus unsurprisingly less frequent in our corpus.

Table 5 — distribution of the subject extractions according to their direction (mean percentage of antepositions and postpositions).

[4] see (Dudewicz & Mishra, 1988)

[5] see (Dudewicz & Mishra, 1988).

[6] This value of the standard deviation, which is greater than the corresponding mean value, shows simply that these distributions do not follows a Gauss distribution.

Corpus	subject anteposition	subject postposition	standard deviation
Individual customers	82.9 %	17.1 %	n.c.
Travel agents	77.3 %	22.7 %	n.c.
Total	**80.6 %**	**19.4 %**	**20.4 %**

All things considered, most of the observed extractions preserve the canonical SVO order (table 6a). Thus, in spite of a frequent use of extractions, spoken French infringe hardly some fundamental ordering constraints.

The inventory of non projective structures supports clearly this observation. Discontinuous extractions are indeed very rare in the Air France corpus (table 6b) : non projective structures, which are very embarrassing for most of parsers or language models, do not concern therefore spoken French.

Table 6 — relative importance of the extractions that follow a canonical SVO order (6a : left) and relative importance of projective extractions (6b : right)

Corpus	% of extractions with SVO order	% of speech turns with SVO order
customers	90.4 %	98.6 %
travel agents	90.2 %	99.0 %
Total	**90.3 %**	**98.7 %**

% of continuous extractions	% of continuous speech turns
97.5 %	99.5 %
98.4 %	99.8 %
97.7 %	**99.6 %**

In conclusion, spoken man-machine dialogue in the ATIS domain seems to be noticeably concerned by a weak word order variability that preserves nevertheless a SVO canonical order, whereas global word order freedom is not really observed.

7. Functions and extracted structures

Extracted structures follow some regularities that should be usefully considered for computational purposes. Table 7 presents for instance the distribution of word order variations according to the extracted construction used. This distribution presents a rather high dispersion. It is however possible to distinguish simple inversion as the most frequent extracted construction (60.6 % of the extractions). This predominance is statistically significant (Student mean test of identical distributions : T = 4.473 ; T(0.01) = 2.600). On the opposite, the difference between dislocations (24.9 %) and presentative structures (among which cleaved sentences ; 13.2 %) is not statistically significant (Student mean test of identical distributions : T = 1.366 ; T_{inv}(1.366) = 0.174).

Table 7 — distribution of the word order variations according to their structure.

Corpus		simple inversions	dislocations	presentatives (cleaved sentences)	other constructions
individual	Mean	60.8 %	24.0 %	13.8%	1.4 %
customers	*(St. Dev.)*	*(27.0 %)*	*(17.3 %)*	*(22.7%)*	*(8.8 %)*
travel agents	Mean	60.2 %	28.3 %	10.5%	1.0 %
	(St. Dev.)	*(36.0 %)*	*(36.1%)*	*(20.6%)*	*(8.1 %)*
Total	Mean	**60.6 %**	**24.9 %**	**13.2 %**	**1.3 %**
	(St. Dev).	*(30.2 %)*	*(25.5 %)*	*(22.1 %)*	*(8.6 %)*

A detailed analysis of these distributions according to the syntactic function of the extracted element provides further conclusions on these structural regularities. Thus, most of subjects and to a lesser extent most of arguments extractions are linguistically marked (dislocations and cleaved structures : table 8).

Table 8 — distribution of subjects and arguments extractions according to their structure.

Corpus	subject extractions		argument extractions	
	inversion	dislocation + presentative	inversion	dislocation + presentative
customers	4.2 %	95.8 %	30.5 %	69.5 %
travel agents	6.1 %	93.9 %	44.4 %	55.6 %
Total	**4.6 %**	**95.4 %**	**32.7 %**	**67.3 %**

This predominance of marked extractions for the argument function is statistically significant (χ^2 test on a not significant predominance : CHI_{AF} = 0.945). On the opposite, adjuncts and sentence complements extractions corresponds almost always to a simple inversion (table 9).

Table 9 — distribution of adjuncts and arguments extractions according to their structure.

Corpus	adjunct extractions		sentence complement extractions	
	inversion	**dislocation + presentative**	**inversion**	**dislocation + presentative**
customers	97.1 %	2.9 %	100.0 %	0.0 %
travel agents	95.3 %	4.7 %	100.0 %	0.0 %
Total	**96.8 %**	**3.2 %**	**100.0 %**	**0.0 %**

Once again, these observations are coherent from a linguistic point of view. On the one hand (subject or argument extraction), cleaved structures or clitics in dislocations compensate partially for an eventual change of the canonical SVO order. On the other hand, adjuncts or sentence complements extraction does not require such marked constructions, since their position is relatively free.

8. Influence of the familiarisation with the task

In the previous sections, we have only considered global observations on the whole Air France corpus. Now, any significant difference between the "customers" and the "travel agents" corpora may show an interesting influence of the familiarisation with the task on word order variations.

The distinction between these two corpora seems a priori relevant. Dialogues are indeed more direct with travel agents, whereas negotiations and reformulatings are noticeably more frequent with individual customers. This observation should be put together with the fact that dialogues with travel agents are shorter than with the other ones. A Wilcoxon-Mann-Withney test[7] shows that this difference is statistically significant (Z = 3.819 ; Z(0.01) = 2.576).

Table 10 — dialogue length according to kind of user (mean number of speech turns per dialogue).

Corpus	mean dialogue length	Standard deviation
individual customers	54.1	33.5
travel agents	42.1	26.4

The familiarisation with the task has therefore a noticeable influence on the dialogue structure. Does this influence concern extractions too ? An exhaustive comparison of the results detailed on tables 2 to 9 suggests that word order variations are independent of this familiarisation. Student mean tests (table 11) show indeed that there is no statistically significant influence on the different features that should characterise word order variations. This observation is obviously interesting for genericity purposes.

Table 11 — Statistical tests (Student mean test) of significance of a feature difference between the "individual customers" and "travel agents" corpora.

Feature	T	T(0.1)	T_{inv}(T)
frequency of occurrence	0.628		0.532

[7] see (Dudewicz & Mishra, 1988)

direction		0.284		0.777
kind of construction	*inversion*	0.212		0.833
	dislocation	0.943		0.348
	presentative	0.715	1.660	0.476
syntactic function	*subject*	0.503		0.616
	argument	0.220		0.827
	adjunct	0.213		0.832
	sentence complement	0.154		0.878
projectivity		0.380		0.705

9. Conclusion : extractions and NLP for man-machine communication

Since they ignore usually most of the syntactic structure of the sentence, current spoken dialogue systems have not been concerned so far by the problem of word order freedom. They circumvent indeed this problem thanks to *ad hoc* approaches that take advantage of the very restricted nature of the considered task. This would not be the case with richer applications or finer dialogue models. As a result, the question of word order freedom will arise soon because of the increasing need[8] of a more detailed language modeling. Now, this corpus analysis provides several lessons on word order freedom that are interesting from a computational point of view.

First of all, the question of discontinuity (global word order freedom), which is very embarrassing for natural language processing, does not concern fortunately spoken French in the ATIS domain. Since discontinuous extractions appear to be very rare, the parsing of these non projective structures does not constitute a relevant problematic for future researches on spoken dialogue systems.

On the opposite, the processing of weak word order freedom should meet an increasing importance as more complex applications will be considered by spoken man-machine communication. The frequent occurrence of extracted constructions in the Air France corpus shows that this question should not be disposed of anymore. Our inventory of several structural regularities (canonical SVO order, specific use of each extracted construction) suggests fortunately that the robust processing of extractions is not an impossible task.

Finally, this corpus analysis does not revealed any influence of the familiarisation with the task on word order variations. This is an interesting result that guarantees to some (restricted) extent the genericity of spoken dialogue systems. This study did not investigate however the independence of word order variations from the application domain. In order to answer this important question, we are currently analysing two additional corpus whose application domain is tourism information. Since the corresponding tasks are clearly less restricted than in the ATIS domain, we hope to obtain interesting conclusions on genericity. Preliminary results suggest that word order variations are rather independent from the application domain, but also that other factors (degree of interactivity for instance) should affect noticeably the frequency of occurrence of spoken extractions (Antoine, 2001).

One aim of this paper was to show the benefit that spoken man-machine communication should obtain from a rigorous analysis of representative corpora. Besides the question of word order variations, we hope that this paper came up — at least partially — to this expectation.

Acknowledgements

The authors are thankful to Agnès Hamon and Valérie Monbet (SABRES Laboratory of Applied Statistics, Vannes, France). They were a great assistance for the achievement of the statistical tests of significance presented in this paper.

[8] For instance, see (Chelba and Jelinek, 2000) for an illustration of the need of strutured language models in speech recognition

References

Antoine J-Y Goulian J 2001 Linguistique de corpus et ingénierie des langues appliquée à la CHM orale : étude des phénomènes d'extraction en français parlé sur deux corpus de dialogue oral finalisé, *TAL*, Hermès Paris France (submitted).

Antoine J-Y Siroux J Caelen J Villaneau J Goulian J Ahafhaf M 2000 Obtaining predictive results with an objective evaluation of spoken dialogue systems : experiments with the DCR assessment paradigm, In *proceedings of the 2nd conference on Language Ressources and Evalaution, LREC'2000*, Athens, Greece.

Blanche-Benveniste C Bilger M Rouget C and van den Eynde K 1990 *Le français parlé : études grammaticales*, CNRS Editions Paris France.

Blanche-Benveniste C 1997 *Approches de la langue parlée en français*, Orphys Paris France.

Chelba C Jelinek F 2000 Structured language modeling, *Computer Speech and Language* 14(4) pp. 283-332, Academic Press London UK.

Covington M 1990 *A dependency parser for variable-order languages*, research Report AI-1990-01, University of Georgia, USA.

Dudewicz E J Mishra S N 1988 *Modern mathematical statistics*, Wiley series in probability and mathematical statistics, John Wiley & Sons New-York, USA.

Gadet F 1992 *Le français populaire*, PUF Paris France.

Hirschman L 1998 Language understanding evaluations : lessons learned from MUC and ATIS. In *proceedings of the 1st conference on Language Ressources and Evalaution, LREC'98*, Granada Spain, pp 117-122.

Holan T, Kubon, Oliva K, Plátek M 2000 On complexity of word order, *TAL* 41(1) pp. 273-300, Hermès Paris.

Hudson R 2000 Discontinuity, *TAL* 41(1) pp. 15-56, Hermès Paris.

Mel'cuk 1988 *Dependency syntax : theory and practice*, State University of New York Press, Albany, USA.

Minker W, Waibel A, Mariani J 1999 *Stochastically based semantic analysis*, Kluwer ac. Amsterdam the Netherlands.

Morel M-A 1989 *Analyse linguistique de corpus*, Publications de la Sorbonne Nouvelle Paris France.

Pollard C, Sag I 1994 *Head-driven Phrase Structure Grammar*, University of Chicago Press, Chicago, USA.

Rambow O, Joshi A 1994 A formal look at dependency grammars and phrase-structure grammars with special considerations of word-order phenomena, In Wanner L. (ed.) *Current issues in Meaning-Text Theory*, Pinter London UK.

Tesnière L 1959 *Elements de syntaxe structurale*, Klincksiek Paris France.